

BioQuant Cluster Users Meeting 2010

BioQuant IT Group

July 29, 2010

Overview

- 1 Overview
- 2 Available Hardware

Overview

- 1 Overview
- 2 Available Hardware
- 3 Software

Overview

- 1 Overview
- 2 Available Hardware
- 3 Software
- 4 Access

Overview

- 1 Overview
- 2 Available Hardware
- 3 Software
- 4 Access
- 5 Available Queues

Overview

- 1 Overview
- 2 Available Hardware
- 3 Software
- 4 Access
- 5 Available Queues
- 6 Resources

Overview

- 1 Overview
- 2 Available Hardware
- 3 Software
- 4 Access
- 5 Available Queues
- 6 Resources
- 7 Job Submission

Overview

- 1 Overview
- 2 Available Hardware
- 3 Software
- 4 Access
- 5 Available Queues
- 6 Resources
- 7 Job Submission
- 8 Monitoring

Overview

- 1 Overview
- 2 Available Hardware
- 3 Software
- 4 Access
- 5 Available Queues
- 6 Resources
- 7 Job Submission
- 8 Monitoring
- 9 Limitations and Policies

Overview

- 1 Overview
- 2 Available Hardware
- 3 Software
- 4 Access
- 5 Available Queues
- 6 Resources
- 7 Job Submission
- 8 Monitoring
- 9 Limitations and Policies
- 10 Future Directions

Overview

- 1 Overview
- 2 Available Hardware
- 3 Software
- 4 Access
- 5 Available Queues
- 6 Resources
- 7 Job Submission
- 8 Monitoring
- 9 Limitations and Policies
- 10 Future Directions
- 11 Epilogue

Overview

- 1 Overview
- 2 Available Hardware
- 3 Software
- 4 Access
- 5 Available Queues
- 6 Resources
- 7 Job Submission
- 8 Monitoring
- 9 Limitations and Policies
- 10 Future Directions
- 11 Epilogue

Overview

- A total of 101 computing nodes are available.
- Two nodes for GPU-based computing
- The nodes are named cln001-int to cln100-int, teslal, teslar.

Overview

- A total of 101 computing nodes are available.
- Two nodes for GPU-based computing
- The nodes are named cln001-int to cln100-int, teslal, teslar.
- Two batch servers in HA-Mode operate the cluster.

Overview

- A total of 101 computing nodes are available.
- Two nodes for GPU-based computing
- The nodes are named cln001-int to cln100-int, teslal, teslar.
- Two batch servers in HA-Mode operate the cluster.
- An installation server takes care of the node installation.

Overview

- A total of 101 computing nodes are available.
- Two nodes for GPU-based computing
- The nodes are named cln001-int to cln100-int, teslal, teslar.
- Two batch servers in HA-Mode operate the cluster.
- An installation server takes care of the node installation.
- The nodes are connected over two dedicated Gigabit networks.

Overview

- A total of 101 computing nodes are available.
- Two nodes for GPU-based computing
- The nodes are named cln001-int to cln100-int, teslal, teslar.
- Two batch servers in HA-Mode operate the cluster.
- An installation server takes care of the node installation.
- The nodes are connected over two dedicated Gigabit networks.

Available Hardware

- 4x Sun Fire x4100 2x Dual-Core AMD Opteron 2218 8GB Memory
- 27x Sun Fire x4100 2x Dual-Core AMD Opteron 2220 8GB Memory
- 1x Sun Fire x4100 2x Dual-Core AMD Opteron 2220 20GB Memory
- 3x Sun Fire x4100 2x Dual-Core AMD Opteron 2220 24GB Memory
- 2x SunFire 4600 8x Quad-Core AMD Opteron 8384 256GB Memory
- 14x IBM xSeries 2x Quad Intel Xeon E5530 16GB Memory
(courtesy of the Schwarz Group)
- 28x IBM xSeries 2x Quad Intel Xeon E5530 20GB Memory (courtesy of the Russell Group)
- 20x Supermicro 2x Quad Intel Xeon E5520 12GB Memory (courtesy of the TIGA Center)
- NVIDIA Tesla S1070 accessible over two frontend nodes

Network Connectivity

The cluster nodes are interconnected via two 1 Gigabit Networks:

Network 1

is a dedicated public network used for file traffic (NFS)

Network 2

is a dedicated private network used for PBS Management and MPI traffic.

Network Connectivity

The cluster nodes are interconnected via two 1 Gigabit Networks:

Network 1

is a dedicated public network used for file traffic (NFS)

Network 2

is a dedicated private network used for PBS Management and MPI traffic.

Software

Batch System

- Batch Server: Torque 2.3.7
- Scheduler: Maui 3.2.6p21

Software

- OS: CentOS 5.4 x86_64 Kernel 2.6.18-164.el5
- GCC: Red Hat 4.1.2-46
- Intel Compiler 11.0
- OpenMPI 1.4.1-gcc and 1.4.1-icc

On selected nodes

The following packages are available on selected nodes:

- Matlab 7.8.0.347 (R2009a) 64-bit
- Mathematica 7
- Maple 13
- cuda 2.3

Software

Batch System

- Batch Server: Torque 2.3.7
- Scheduler: Maui 3.2.6p21

Software

- OS: CentOS 5.4 x86_64 Kernel 2.6.18-164.el5
- GCC: Red Hat 4.1.2-46
- Intel Compiler 11.0
- OpenMPI 1.4.1-gcc and 1.4.1-icc

On selected nodes

The following packages are available on selected nodes:

- Matlab 7.8.0.347 (R2009a) 64-bit
- Mathematica 7
- Maple 13
- cuda 2.3

Access

Prerequisites

The only prerequisite for accessing the cluster is a BioQuant account.

Submit Hosts

Jobs can be submitted to the cluster from our Linux Applications Servers:
`app1{1|2|9|10}`

Access

Prerequisites

The only prerequisite for accessing the cluster is a BioQuant account.

Submit Hosts

Jobs can be submitted to the cluster from our Linux Applications Servers:
`app1{1|2|9|10}`

Queues

`short`

```
{default|max}.cpus=01:00:00
```

`research`

```
{default|max}.cpus=05:00:00
```

Queues

short

```
{default|max}.cput=01:00:00
```

research

```
{default|max}.cput=05:00:00
```

quick

```
{default|max}.cput=00:15:00
```

batch

```
default.cput=168:00:00; default.walltime=336:00:00 (max. 30 Running Jobs/User)
```

Queues

short

```
{default|max}.cput=01:00:00
```

research

```
{default|max}.cput=05:00:00
```

quick

```
{default|max}.cput=00:15:00
```

batch

```
default.cput=168:00:00; default.walltime=336:00:00 (max. 30 Running Jobs/User)
```

Warning: you cannot submit into those queues directly. The right queue is determined by PBS using the requested time for the job; it defaults to batch

Queues

short

```
{default|max}.cput=01:00:00
```

research

```
{default|max}.cput=05:00:00
```

quick

```
{default|max}.cput=00:15:00
```

batch

```
default.cput=168:00:00; default.walltime=336:00:00 (max. 30 Running Jobs/User)
```

Warning: you cannot submit into those queues directly. The right queue is determined by PBS using the requested time for the job; it defaults to batch

cuda

special queue for jobs that run on the Tesla machines; use `-q cuda` in the `qsub` command line to submit to this queue.

Queues

short

```
{default|max}.cput=01:00:00
```

research

```
{default|max}.cput=05:00:00
```

quick

```
{default|max}.cput=00:15:00
```

batch

```
default.cput=168:00:00; default.walltime=336:00:00 (max. 30 Running Jobs/User)
```

Warning: you cannot submit into those queues directly. The right queue is determined by PBS using the requested time for the job; it defaults to batch

cuda

special queue for jobs that run on the Tesla machines; use `-q cuda` in the `qsub` command line to submit to this queue.

Resources

Some PBS Integrated Resources

- ppn: # processors per node
- mem: # total amount of memory
- cput: # Maximum amount of CPU time used by all processes in the job (seconds, or [[HH:]MM:]SS)
- walltime: # Maximum amount of real time during which the job can be in the running state

BioQuant HPC Specific Resources

- matlab # a node with a Matlab installation
- math # a node with a Mathematica installation
- maple # a node with a Maple installation
- xeon # a node with an Intel Xeon Processor
- opteron # a node with an AMD Opteron Processor

Resources

Some PBS Integrated Resources

- ppn: # processors per node
- mem: # total amount of memory
- cput: # Maximum amount of CPU time used by all processes in the job (seconds, or [[HH:]MM:]SS)
- walltime: # Maximum amount of real time during which the job can be in the running state

BioQuant HPC Specific Resources

- matlab # a node with a Matlab installation
- math # a node with a Mathematica installation
- maple # a node with a Maple installation
- xeon # a node with an Intel Xeon Processor
- opteron # a node with an AMD Opteron Processor

Job Submission I: Interactive Sessions

Simple Interactive Session

```
qsub -I
```

Interactive Session with X forwarding

```
qsub -X -I
```


Job Submission I: Interactive Sessions

Simple Interactive Session

```
qsub -I
```

Interactive Session with X forwarding

```
qsub -X -I
```

Node-exclusive Interactive Session

```
qsub -I [-X] -W x=NACCESSPOLICY=SINGLEJOB
```

Job Submission I: Interactive Sessions

Simple Interactive Session

```
qsub -I
```

Interactive Session with X forwarding

```
qsub -X -I
```

Node-exclusive Interactive Session

```
qsub -I [-X] -W x=NACCESSPOLICY=SINGLEJOB
```

Job Submission II: More Interactive Sessions

- `qsub -I -l ncpus=4 # ask for four tasks (processes)`

Job Submission II: More Interactive Sessions

- `qsub -I -l ncpus=4 # ask for four tasks (processes)`
- `qsub -I -l nodes=4:ppn=4 # ask for four nodes with four processes on each node.`

Job Submission II: More Interactive Sessions

- `qsub -I -l ncpus=4 # ask for four tasks (processes)`
- `qsub -I -l nodes=4:ppn=4 # ask for four nodes with four processes on each node.`
- `qsub -I -l nodes=1:matlab # ask for a node that has the matlab feature.`

Job Submission II: More Interactive Sessions

- `qsub -I -l ncpus=4 # ask for four tasks (processes)`
- `qsub -I -l nodes=4:ppn=4 # ask for four nodes with four processes on each node.`
- `qsub -I -l nodes=1:matlab # ask for a node that has the matlab feature.`
- `qsub -I -l mem=200gb # ask for a node with 200gb of memory`

Job Submission II: More Interactive Sessions

- `qsub -I -l ncpus=4 # ask for four tasks (processes)`
- `qsub -I -l nodes=4:ppn=4 # ask for four nodes with four processes on each node.`
- `qsub -I -l nodes=1:matlab # ask for a node that has the matlab feature.`
- `qsub -I -l mem=200gb # ask for a node with 200gb of memory`
- `qsub -I -l nodes=2,mem=200gb,cput=00:15:00 # ask for two nodes with a total of 200gb of memory for 15 minutes of processing time`

Job Submission II: More Interactive Sessions

- `qsub -I -l ncpus=4 # ask for four tasks (processes)`
- `qsub -I -l nodes=4:ppn=4 # ask for four nodes with four processes on each node.`
- `qsub -I -l nodes=1:matlab # ask for a node that has the matlab feature.`
- `qsub -I -l mem=200gb # ask for a node with 200gb of memory`
- `qsub -I -l nodes=2,mem=200gb,cput=00:15:00 # ask for two nodes with a total of 200gb of memory for 15 minutes of processing time`
- `qsub -I -q cuda # ask that the job be submitted to the cuda queue (will be executed on a Tesla machine; see above: Hardware)`

Job Submission II: More Interactive Sessions

- `qsub -I -l ncpus=4 # ask for four tasks (processes)`
- `qsub -I -l nodes=4:ppn=4 # ask for four nodes with four processes on each node.`
- `qsub -I -l nodes=1:matlab # ask for a node that has the matlab feature.`
- `qsub -I -l mem=200gb # ask for a node with 200gb of memory`
- `qsub -I -l nodes=2,mem=200gb,cput=00:15:00 # ask for two nodes with a total of 200gb of memory for 15 minutes of processing time`
- `qsub -I -q cuda # ask that the job be submitted to the cuda queue (will be executed on a Tesla machine; see above: Hardware)`

An example PBS Script

Listing 1: Sample PBS Script

```
1  #!/bin/sh
2  # Following that, we can use the first PBS directive to set the name of our job:
3  # Job Name (this line is a comment and won't be processed)
4  #PBS -N my_pbs_job
5  # Now we can ask for some resources, e.g. a total time of 15 hours and one node
6  # with 4 processors:
7  #PBS -l walltime=15:00:00
8  #PBS -l nodes=1:ppn=4
9  # A most important directive is the one which ensures that the Linux
10 # environment for the job is the same as the one we're working in:
11 #PBS -V
12
13 # Then we define an alternative e-mail address where PBS messages will be sent
14 # to (default is local mail)...
15 #PBS -M joe.doe@bioquant.uni-heidelberg.de
16 # ...and specify that an e-mail should be sent to the user when the job
17 # begins (b), ends (e) or aborts (a)
18 #PBS -m bea
19 # Define a file where stderr will be redirected to
20 #PBS -e my_pbs_job.err
21 # Define a file where stdout will be redirected to
22 #PBS -o my_pbs_job.log
23 # finally we let our job run as usual
24 ./my-job
25 # or if it is an MPI job
26 mpirun ./my_mpi-job
```

pbsdsh

pbsdsh

The PBS distributed shell - reads the file `$PBS_O_NODELIST` and executes a command on all reserved nodes.

pbsdsh Control

If the command is itself a script, we can control executions through the `$PBS_VNODENUM` variable; we can either send a different executable to each processor or provide each instance of the same executable with different input.

A simple example follows:

Listing 2: pbsdsh Call

```
—bash—3.2$ pbsdsh —s my_script.sh
```

Listing 4: pbsdsh Call

```
—bash—3.2$ pbsdsh —s my_script.sh
```

Listing 3: my_script.sh

```
#!/bin/sh
$PBS_O_WORKDIR/myprog.$PBS_VNODENUM
```

Listing 5: my_script.sh

```
#!/bin/sh
$PBS_O_WORKDIR/myprog<some_data.$PBS_VNODENUM
```

Monitoring Tools

- `qstat`
- `Ganglia`

Monitoring Tools

- `qstat`
- `Ganglia`

qstat

qstat

displays the information the PBS Batch System has for a job; among other things, information about reserved resources, running time, current state can be fetched.

Job States

Q	Queued (Waiting for resources)
R	Running
B	Blocked
H	Halted through qhold
C	Complete (Finished or terminated)

```

a2 :
File Edit View Scrollback Bookmarks Settings Help
bq_gnikolis@appl2:~$ qstat 511187
Job id          Name          User          Time Use S Queue
-----
511187.cln035  STDIN        bq_gnikolis  00:00:00 R batch
bq_gnikolis@appl2:~$
  
```

qstat

qstat

displays the information the PBS Batch System has for a job; among other things, information about reserved resources, running time, current state can be fetched.

Job States

- Q Queued (Waiting for resources)
- R Running
- B Blocked
- H Halted through qhold
- C Complete (Finished or terminated)

```

a2 :
File Edit View Scrollback Bookmarks Settings Help
bq_gnikolis@appl2:~$ qstat 511187
Job id          Name          User          Time Use S Queue
-----
511187.cln035  STDIN        bq_gnikolis  00:00:00 R batch
bq_gnikolis@appl2:~$
  
```

Some qstat Options

Here are some useful options for the qsub command:

- f < *JOBID* > Specifies that a full status display be written to standard out.
- n < *JOBID* > In addition to the basic information, nodes allocated to a job are listed.
- q Specifies that the request is for queue status
- s batch Displays jobs in the batch queue.

Omitting < *JOBID* >

in the first two commands above causes qstat to display the relevant information for all jobs.

Ganglia

Ganglia

Ganglia is an open source monitoring package (not only) for clusters.

Graphs

Apart from information about the current cpu, memory and network load, Ganglia provides intuitive graphs for each node and for the defined partitions of the cluster.

Ganglia

Ganglia

Ganglia is an open source monitoring package (not only) for clusters.

Graphs

Apart from information about the current cpu, memory and network load, Ganglia provides intuitive graphs for each node and for the defined partitions of the cluster.

Web-Interface

The Web Interface of Ganglia is accessible via: <http://c1n035/ganglia>

Ganglia

Ganglia

Ganglia is an open source monitoring package (not only) for clusters.

Graphs

Apart from information about the current cpu, memory and network load, Ganglia provides intuitive graphs for each node and for the defined partitions of the cluster.

Web-Interface

The Web Interface of Ganglia is accessible via: <http://c1n035/ganglia>

Access

The Web Interface is accessible only from within the BioQuant networks.

Ganglia

Ganglia

Ganglia is an open source monitoring package (not only) for clusters.

Graphs

Apart from information about the current cpu, memory and network load, Ganglia provides intuitive graphs for each node and for the defined partitions of the cluster.

Web-Interface

The Web Interface of Ganglia is accessible via: <http://c1n035/ganglia>

Access

The Web Interface is accessible only from within the BioQuant networks.

Ganglia View

Ganglia: BioQuant HPC Grid Report - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://cln035/ganglia/?r=hour&s=by%2520hosts%2520up&c=

heise online News Linux-Magazin BIOQUANT Google Mail - Inbox The Python Stand... MySQL :: MySQL 5... 20 Linux System M... Sun DS Contents Bit Calculator - Co...

Ganglia: BioQuant HPC Grid Rep...

Ganglia BioQuant HPC Grid Report for Tue, 06 Jul 2010 11:08:39 +0200

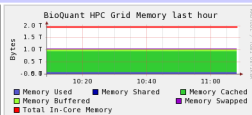
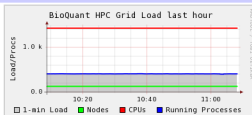
Last Sorted

BioQuant HPC Grid >

BioQuant HPC Grid (6 sources) (tree view)

CPU's Total: **1424**
 Hosts up: **128**
 Hosts down: **0**

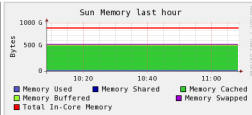
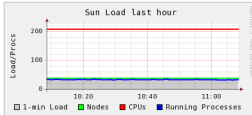
Avg Load (15, 5, 1m):
 29%, 29%, 29%
 Localtime:
 2010-07-06 11:08



Sun (physical view)

CPU's Total: **208**
 Hosts up: **38**
 Hosts down: **0**

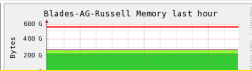
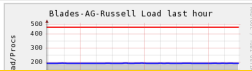
Avg Load (15, 5, 1m):
 18%, 16%, 16%
 Localtime:
 2010-07-06 11:08



Blades-AG-Russell (physical view)

CPU's Total: **448**
 Hosts up: **28**
 Hosts down: **0**

Avg Load (15, 5, 1m):



Limitations

max_user_run

For the batch queue a limit of max. 30 running jobs/user is in place. This does not in principle prevent a user from filling up the cluster but has been so far proven useful in keeping things unter control. For the other queues no such constraint is defined.

ssh and rsh

Both remote execution tools are disabled on the cluster

Limitations

max_user_run

For the batch queue a limit of max. 30 running jobs/user is in place. This does not in principle prevent a user from filling up the cluster but has been so far proven useful in keeping things under control. For the other queues no such constraint is defined.

ssh and rsh

Both remote execution tools are disabled on the cluster

Maui: #define MMAX_JOB 4096

Due to this macro in `msched.h`, the Maui Scheduler cannot hold more than 4096 jobs at a time. This number includes running, queued, held and blocked jobs. Once this limit is exceeded, new jobs are simply discarded. Keep that in mind when you let a script submitting jobs into the cluster :-)

Limitations

`max_user_run`

For the batch queue a limit of max. 30 running jobs/user is in place. This does not in principle prevent a user from filling up the cluster but has been so far proven useful in keeping things under control. For the other queues no such constraint is defined.

`ssh` and `rsh`

Both remote execution tools are disabled on the cluster

Maui: `#define MMAX_JOB 4096`

Due to this macro in `msched.h`, the Maui Scheduler cannot hold more than 4096 jobs at a time. This number includes running, queued, held and blocked jobs. Once this limit is exceeded, new jobs are simply discarded. Keep that in mind when you let a script submitting jobs into the cluster :-)

Policies

Interactive Sessions

Interactive sessions on the cluster should be closed when they are no longer needed.

Multi-threaded and memory intensive jobs

These kind of jobs should be submitted with the SINGLEJOB flag in order not to interfere with other jobs which are eventually running simultaneously on the same node.

Policies

Interactive Sessions

Interactive sessions on the cluster should be closed when they are no longer needed.

Multi-threaded and memory intensive jobs

These kind of jobs should be submitted with the SINGLEJOB flag in order not to interfere with other jobs which are eventually running simultaneously on the same node.

Node Reservation

Never reserve more nodes than you need / can use. The only ways to run a job on multiple nodes is either via `pbsdsh` or via an MPI executable. All other jobs are confined to one node.

Policies

Interactive Sessions

Interactive sessions on the cluster should be closed when they are no longer needed.

Multi-threaded and memory intensive jobs

These kind of jobs should be submitted with the SINGLEJOB flag in order not to interfere with other jobs which are eventually running simultaneously on the same node.

Node Reservation

Never reserve more nodes than you need / can use. The only ways to run a job on multiple nodes is either via `pbsdsh` or via an MPI executable. All other jobs are confined to one node.

Future Directions

- 10Gb Network on specific nodes for MPI; do we need it? User feedback.
- Direct Access to LSDF from the Cluster
- Renewal of the BIOMS-Cluster

Epilogue

That's it!

Thank you for you attention!

Contact Information

Send questions, inquiries and complaints to:
cluster@bioquant.uni-heidelberg.de

Mailing List

BQ-CLUSTER@bioquant.uni-heidelberg.de

For news, developments and communication about the BioQuant Cluster and its usage.